Nicolas CHATRON ORCID iD: 0000-0003-0538-0981

Cedric Le Caignec ORCID iD: 0000-0002-0598-653X

Identification of mobile retrocopies during genetic testing: consequences for routine diagnosis

Authors' full name

Nicolas Chatron [1,2], Kevin Cassinari [3], Olivier Quenez [3], Stéphanie Baert-Desurmont [4], Claire Bardel [5,6], Marie-Pierre Buisine [7], Eduardo Calpena [8],Yline Capri [9], Jordi Corominas Galbany [10], Flavie Diguet [1,2], Patrick Edery [1,2] , Bertrand Isidor [11], Audrey Labalme[1], Cedric Le Caignec [11,12], Jonathan Lévy [13], François Lecoquierre [4], Pierre Lindenbaum [14,15], Olivier Pichon [11], Pierre-Antoine Rollat-Farnier [1,5], Thomas Simonet [16,17], Pascale Saugier-Veber [4], Anne-Claude Tabet [13,18], Annick Toutain [19,20], Andrew O. M. Wilkie [8], Gaetan Lesca [1,2], Damien Sanlaville [1,2], Gaël Nicolas [3], Caroline Schluth-Bolard [1,2]

Authors' primary affiliations

1. Service de génétique, Hospices Civils de Lyon, Lyon, France
2. Equipe GENDEV, CRNL, INSERM U1028, CNRS UMR5292, UCBL1, Lyon, France
3. Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics and CNR-MAJ, F 76000, Normandy Center for Genomic and Personalized Medicine, Rouen, France

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23845.

4. Normandie Univ, UNIROUEN, Inserm U1245 and Rouen University Hospital, Department of Genetics, F 76000, Normandy Center for Genomic and Personalized Medicine, Rouen, France
5. Cellule bioinformatique de la plateforme de séquençage NGS du CHU de Lyon, Groupement Hospitalier Est, Lyon, France
6. Service de biostatistique bioinformatique, HCL, Lyon, France
7. Inserm UMR-S 1172, JPA Research Center, Lille University, and Department of Biochemistry and Molecular Biology, Lille University Hospital, Lille, France.
8. Clinical Genetics Group, MRC Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK
9. UF de génétique clinique, Département de génétique, Hôpital Universitaire Robert Debré, AP-HP, Paris, France
10. Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands
11. CHU Nantes, Service de Génétique Médicale, Nantes, France
12. Université de Nantes, Nantes, France
13. UF de cytogénétique, Département de génétique, Hôpital Universitaire Robert Debré, AP-HP, Paris, France
14. INSERM, UMR_S1087, l'institut du thorax, Nantes, France
15. CNRS, UMR 6291, Nantes, France
16. Centre de Biotechnologie Cellulaire, Hospices Civils de Lyon, Lyon, France
17. Nerve-Muscle Interactions Team, Institut NeuroMyoGène CNRS UMR 5310 - INSERM U1217 - Université Claude Bernard Lyon 1

18. Unité Génétique Humaine et Fonction Cognitive, Département de neurosciences, Institut Pasteur, Paris, France
19. Service de Génétique, Hôpital Bretonneau, CHU, Tours, France 20
20. UMR 1253, iBrain, Université de Tours, Inserm, Tours, France

Corresponding author

Dr Nicolas Chatron

Service de génétique, Centre de Biologie Pathologie Est 2ème étage

Groupement Hospitalier Est, Hospices Civils de Lyon

59 Boulevard Pinel, 69677 Bron CEDEX, France

Tel : +33472129640

Mail : nicolas.chatron@chu-lyon.fr

Conflict of interest

The authors declare no conflict of interest.

Abstract

Human retrocopies, i.e. mRNA transcripts benefitting from the LINE-1 machinery for retrotransposition, may have specific consequences for genomic testing. NGS techniques allow the detection of such mobile elements but they may be misinterpreted as genomic duplications or be totally overlooked. We report eight observations of retrocopies detected during diagnostic NGS analyses of targeted gene panels, exome, or genome sequencing. For seven cases, while an exons-only copy number gain was called, read alignment inspection revealed a depth of coverage shift at every exon-intron junction where indels were also systematically called. Moreover, aberrant chimeric read pairs spanned entire introns or were paired with another locus for terminal exons. The 8th retrocopy was present in the reference genome and thus

showed a normal NGS profile. We emphasize the existence of retrocopies and strategies to accurately detect them at a glance during genetic testing and discuss pitfalls for genetic testing.

Keywords

Retrocopies, Genome mobility, copy number gain, genetic testing pitfalls

Introduction

Since Barbara McClintock's demonstration using corn (McClintock, 1951), genomic mobility has largely been studied and related to evolutionary processes and genetic diseases. While DNA transposons are inactive in humans (Lander et al., 2001), DNA retrotransposons use a RNA step for mobilization. They have been separated in two groups depending on whether they contain a retrovirus-derived sequence (human endogenous retroviruses (HERVs)) or not. RNA non-ERV elements are the most abundant class covering together one third of the total human reference genome sequence (Lander et al., 2001). Four different types of these can be distinguished (Feschotte and Pritham, 2007). LINE-1 (long interspersed element 1) are the only autonomous elements capable of retrotranscription and insertion; the other three types Alu, SINE-R-VNTR-Alu (SVA) and retrocopies require the LINE-1 machinery for mobility (Esnault et al., 2000; Dewannieux et al., 2003; Hancks et al., 2012; Raiz et al., 2012). Retrocopies (also referenced as retrogene, retroCNV, retroduplication, processed pseudogenes or retrotransposed pseudogenes in the literature) are mRNAs that have been reverse transcribed and inserted in the genome. Studying primate genomes, 127 retrocopies out of 8000 present in the reference human genome could be demonstrated as human-specific (Zhang et al., 2003; Navarro and Galante, 2015). While LINE-1 mediated events are estimated to occur in 1 of 21 meioses for Alu

(Xing et al., 2009), 1 of 95-270 for LINE-1 (Xing et al., 2009; Ewing and Kazazian, 2010) and 1 of 916 for SVA (Xing et al., 2009), retrocopy events are the rarest, estimated at 1 per 6256 meioses in the literature (Ewing et al., 2013).

The first description of a pathogenic mobile element was the insertion of a LINE-1 sequence within *F8* (Kazazian et al., 1988) responsible for haemophilia A and, since then, more than 130 similar events (Kazazian and Moran, 2017) have been detected in diseases genes associated such as Duchenne muscular dystrophy (Smith et al., 2011), beta-thalassemia trait (Lanikova et al., 2013), and haemophilia B (Mukherjee et al., 2004) among many examples. With the spread of genome sequencing-based testing, such observations are likely to become more frequent. Beyond structural variant calling, data interpretation requires expertise in genome biology and the ability to recognize such phenomena. Herein, we present eight observations of retrocopies identified during genetic interpretation of NGS data with diverse techniques and diverse clinical situations.

We detail potential pitfalls for genetic analysis and discuss their possible implications in human diseases.

Material and Methods

All observations of retrocopies from Lyon, Rouen and Nantes University hospitals made since 2016 were included. Patients received gene panel (P1-P3), exome (P4, P8) or genome (P5-P7) sequencing after providing informed written consent. Testing indications and individual methods are summarized in Table 1. All techniques used paired-end sequencing either on an Illumina platform (Illumina Inc., San Diego, CA, USA) (Patients 1-4, 6-8) or BGI technology (Shenzhen, China) (Patient 5).

All sequencing tests used BWA-MEM v.0.7 (Li and Durbin, 2009) for genomic alignment over the GRCh37 assembly. Single nucleotide variant calling was performed using GATK Haplotype Caller v3.8 (P1-P5, P8) while CANOES (Backenroth et al., 2014) (P1, P2, P4), DeCovA (Dimassi et al., 2015) (P3) and ERDS (Zhu et al., 2012) (P5-P7) were used for CNV calling. Finally, structural variants were called using either GRIDSS v1.4.1 (Cameron et al., 2017) for targeted sequencing (P1, P4) or BreakDancer v1.4.5 (Chen et al., 2009) for genome data (P5-P7). IGV 2.3 was used for data visualization (Robinson et al., 2011). Aberrant junction fragments were confirmed using classic PCR amplification (P1-P4, P6). For patient 1, QMPSF was performed using custom fluorescent primers, with the protocol previously described (Charbonnier et al., 2000). Moreover, a commercial kit was used for Multiplex Ligation-dependent Probe Amplification (SALSA MLPA probemix P158-C2 Juvenile Polyposis (JPS), MRC-Holland, Amsterdam, The Netherlands). For Patient 5, droplet digital PCR was performed on the QX200 ddPCR platform (Bio-Rad, Hercules, CA, USA) using hydrolysis probes (TaqMan probes) on target (FAM tag) and *HMBS* reference gene (VIC tag). All primers are available upon request.

Results

Patients 1 to 7 presented a similar profile regarding Single Nucleotide, Copy Number and Structural Variant (SNV, CNV, SV) calling. CNV calling showed an exons-only duplication of genes *SMAD4*, *CCDC88C, DKC1*, *MTMR2, TYRO3, RCBTB1* and *C20orf27* respectively for patient 1 to patient 7 (Table 1). In each case, detailed depth of coverage analysis showed an immediate shift at every exon-intron junction for a specific transcript. Structural variant analysis revealed clustered calls within the duplicated gene and apparent translocations linking both extremities of this same gene to another locus. Finally, SNV calling systematically called large indels at every splice

acceptor and donor sites of a specific transcript. Alignment visualization revealed that indels were exclusively supported by soft-clipped reads (i.e., reads containing 2 concatenated DNA sequences mapping to 2 exon extremities with aberrant exon-exon junctions) paired with reads aligned to the preceding or following exon instead of flanking intronic sequence (Figure 1 for patients 3 and 5, supplemental figure S1 for other patients). Patient's 8 *RPL13* retrocopy was not detected during the initial analysis of NGS data. It could only be identified during the cDNA study of an intronic single nucleotide variant identified in the exome. Several unexpected single nucleotide variants were observed on the Sanger sequence of this RT-PCR product while absent from exome results. Further analysis showed that this sequence could perfectly align to an unknown *RPL13* retrocopy present in the reference genome on chromosome 19 (chr19:56216980-56217800 [hg38]) (supplemental figure S1). The retrocopy had been preferentially amplified and sequenced after inappropriate DNase treatment.

None of these rearrangements directly disrupted a coding sequence or a Topologically Associated Domain (TAD) boundary. While 6 out of 8 insertions occurred in another gene, none was likely to produce a fusion transcript regarding transcription strand or translation phase but exonization could not be ruled out. Events detected for patients 2, 3 and 6 were inherited from a healthy parent and they were considered as likely benign. For patient 5, the *TYRO3* retrocopy was initially detected as a candidate *de novo* event during genome analysis of a sporadic early onset Alzheimer's disease patient – unaffected parents trio, but detailed data inspection and confirmation techniques showed that both parents were heterozygous carriers of the retrocopy and the proband carried it in a homozygous state (Figure 1). In both parents, the insertion occurred within intron 2 of *ENOX1* on chromosome 13 at the same position. In addition, the presence of two SNVs phased to this retrocopy favored a remote non-

functional event spread within the general population. Reviewing our in house exome and genome data, we found this *TYRO3* retrocopy to be a common event among controls, also previously described in the literature (Schrider et al., 2013; MacDonald et al., 2014), confirming its likely benign nature. The patient 1 *SMAD4* retrotransposition within intron 17 of *SCAI* has also been demonstrated to be a rare likely benign event with a frequency of 0.25% in European populations (Millson et al., 2015; Watson et al., 2017). The patient 8 *RPL13* retrocopy is present in the reference genome and could not even be considered as a variant. Finally, looking at chimeric reads spanning exon-exon junctions in the 1000G project dataset of 2535 individual genomes, *MTMR2* and *RCBTB1* retrocopies could also be suspected in 6 and 3 patients respectively.

Discussion

Through these eight observations we aim to emphasize potential consequences of genome mobility on genome biology and genetic testing. We demonstrate that paired-end sequencing standard pipelines easily highlight retrocopies. Their frequency might, however, be underestimated in routine genetic testing, because they can be either overlooked in NGS standard analyses or undetected by other molecular genetics technologies. While LINE-1, Alu and SVAs insertions require dedicated bioinformatics pipelines (Kvikstad et al., 2018), retrocopies present a recognizable footprint on sequencing results characterized by the association between clustered indel calls at exon-intron junctions, a unique exons-only copy number gain, and aberrant read pairs spanning introns on the one hand and between the parental locus and the insertion site on the other hand.

The clinical consequences regarding pathogenicity mostly depend on the insertion site. Indeed, as for other mobile elements, retrocopies could disrupt a gene and cause haploinsufficiency but considering their coding origin, they could also form a fusion transcript with unpredictable consequences, as exemplified by HIV resistance observed in owl monkeys after the insertion of a *Cyclophilin A* cDNA into the *TRIM5* gene (Sayah et al., 2004). Retrocopies could also disturb physiological chromatin interactions and have consequences on neighboring genes' expression. Additionally, because of their coding capacities, retrocopies could theoretically still be transcribed and translated. Retrocopies have been distinguished in retrogenes and retropseudogenes based on their transcribed residual activity (Kaessmann et al., 2009). This requires fortuitous coupling between the insertion site and a favorable regulatory genomic context. An additional copy of the transcript could then be disease-causing either because of a dose effect or through an ectopic expression. Tissue-specific expression of transcribed retrocopies has indeed been demonstrated to be closely related to the insertion site and not to their parental gene (Navarro and Galante, 2015). Finally, if the insertion occurs on the opposite strand compared to the original gene, a siRNA could be produced leading to paradoxical haploinsufficiency (Guo et al., 2009). While several examples of pathogenic retrocopies exist in animals (Sayah et al., 2004; Parker et al., 2009), to our knowledge, a single case of human pathogenic retrocopy has been reported in the literature. De Boer et al. reported a patient with chronic granulomatous due to an abnormal splicing of *CYBB* caused by the partial exonisation of a *TMF1* retrocopy (de Boer et al., 2014). Here, we highlight the fact that retrocopies detected during NGS analyses may also be interpreted as benign, even within gene panels focusing on a few genes with a high *a priori* probability to be disease-causing.

Beyond these clinical aspects, retrocopies are also important to be recognized because of their misleading consequences on different molecular genetics techniques. Looking at SNVs, as seen here using short-read sequencing, aberrant splice site indels are called and will not be confirmed using Sanger sequencing. To address this, using RNA-Seq aligner as an add-on to standard pipelines has been proposed (Watson et al., 2017). Furthermore, SNVs phased to the retrocopy could yield false positive exonic SNV calls on the parent gene especially if PCR primers were selected in exonic region (large exons). Regarding CNVs, using intronic or exonic-intronic qPCR primers, the apparent duplication would not be confirmed either. On the contrary, retrocopies would be seen as entire duplications using MLPA but could mask genic deletions and provoke false negative results as has already been described for *SMAD4* (Millson et al., 2015). Because of backbone design and intronic probes, retrocopies are likely to be missed using cytogenetic microarray as has been the case for patients 1 and 2. Moreover, aberrant read pairs spanning introns and at the insertion site are responsible for structural variant calling. These can be misinterpreted as a chromosomal translocation as has already been the case for the *TYRO3* polymorphic retrocopy misinterpreted as a chr 15 – chr 13 translocation (Brastianos et al., 2013). Finally, the presence of retrocopies also justifies the use of DNase treatment of RNAs for RT-PCR assays as they would mimic RNA and introduce false results as has been the case for Patient 8 (Menon et al., 1991; Mutimer et al., 1998). Interpretation of RNA-Seq data should also be cautious for retrocopies present in the reference genome, as reads could be aligned ambiguously between the retrocopy and its parental transcript. Bioinformatics pipelines should take this risk into account.

We propose here a description that can enable the identification of retrocopies at a glance and thus (i) prevent from false interpretation, (ii) warn about potential false

positives and false negatives (depending on technique) due to such mechanism and (iii) help to figure out potential mechanisms of pathogenicity.

Acknowledgements

We are thankful to Joris Veltman, Lisenka Vissers, Nathalie Drouot, Christian Gilissen and Reza Maroofian. G.N. acknowledges Fondation Bettencourt-Schueller, Fondation Philippe Chatrier, Fondation Charles Nicolle, and Association Cerveau Progrès. This work was supported by the NIHR Oxford Biomedical Research Centre (AOMW). Patient's 7 whole genome sequencing was funded by independent research commissioned by the Health Innovation Challenge Fund (R6-388 / WT 100127), a parallel funding partnership between Wellcome and the Department of Health (Principal Investigator: Jenny C Taylor). The views expressed in this publication are those of the authors and not necessarily those of Wellcome or the Department of Health.

References

Backenroth D, Homsy J, Murillo LR, Glessner J, Lin E, Brueckner M, Lifton R, Goldmuntz E, Chung WK, Shen Y. 2014. CANOES: detecting rare copy number variants from whole exome sequencing data. Nucleic Acids Res 42:e97–e97.

Boer M de, Leeuwen K van, Geissler J, Weemaes CM, Berg TK van den, Kuijpers TW, Warris A, Roos D. 2014. Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. Hum Mutat 35:486–96.

Brastianos PK, Horowitz PM, Santagata S, Jones RT, McKenna A, Getz G, Ligon KL, Palescandolo E, Hummelen P Van, Ducar MD, Raza A, Sunkavalli A, et al. 2013. Genomic sequencing of meningiomas identifies oncogenic SMO and AKT1 mutations. Nat Genet 45:285–9.

Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT. 2017. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27:2050–2060.

Charbonnier F, Raux G, Wang Q, Drouot N, Cordier F, Limacher JM, Saurin JC, Puisieux A, Olschwang S, Frebourg T. 2000. Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments.

Cancer Res 60:2760–3.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677–81.

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet 35:41–48.

Dimassi S, Simonet T, Labalme A, Boutry-Kryza N, Campan-Fournier A, Lamy R, Bardel C, Elsensohn M-H, Roucher-Boulez F, Chatron N, Putoux A, Bellescize J de, et al. 2015. Comparison of two next-generation sequencing kits for diagnosis of epileptic disorders with a user-friendly tool for displaying gene coverage, DeCovA. Appl Transl genomics 7:19–25.

Esnault C, Maestre J, Heidmann T. 2000. Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363–367.

Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol 14:R22.

Ewing AD, Kazazian HH. 2010. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20:1262–70.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–68.

Guo X, Zhang Z, Gerstein MB, Zheng D. 2009. Small RNAs Originated from Pseudogenes: cis- or trans-Acting? PLoS Comput Biol 5:e1000449.

Hancks DC, Mandal PK, Cheung LE, Kazazian HH. 2012. The Minimal Active Human SVA Retrotransposon Requires Only the 5’-Hexamer and Alu-Like Domains. Mol Cell Biol 32:4718–4726.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication:

mechanistic and evolutionary insights. Nat Rev Genet 10:19–31.

Kazazian HH, Moran J V. 2017. Mobile DNA in Health and Disease. N Engl J Med 377:361–370.

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164–166.

Kvikstad EM, Piazza P, Taylor JC, Lunter G. 2018. A high throughput screen for active human transposable elements. BMC Genomics 19:115.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lanikova L, Kucerova J, Indrak K, Divoka M, Issa J-P, Papayannopoulou T, Prchal JT, Divoky V. 2013. β-Thalassemia Due to Intronic LINE-1 Insertion in the *β-Globin* Gene (*HBB*): Molecular Mechanisms Underlying Reduced Transcript Levels of the *β-Globin* $_{L1}$ Allele. Hum Mutat 34:1361–1365.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42:D986-92.

McClintock B. 1951. Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 16:13–47.

Menon RS, Chang YF, St Clair J, Ham RG. 1991. RT-PCR artifacts from processed pseudogenes. PCR Methods Appl 1:70–1.

Millson A, Lewis T, Pesaran T, Salvador D, Gillespie K, Gau C-L, Pont-Kingdon G, Lyon E, Bayrak-Toydemir P. 2015. Processed Pseudogene Confounding Deletion/Duplication Assays for SMAD4. J Mol Diagnostics 17:576–582.

Mukherjee S, Mukhopadhyay A, Banerjee D, Chandak GR, Ray K. 2004. Molecular pathology of haemophilia B: identification of five novel mutations including a LINE 1 insertion in Indian patients. Haemophilia 10:259–263.

Mutimer H, Deacon N, Crowe S, Sonza S. 1998. Pitfalls of Processed Pseudogenes in RT-PCR. Biotechniques 585–588.

Navarro FCP, Galante PAF. 2015. A Genome-Wide Landscape of Retrocopies in Primate Genomes. Genome Biol Evol 7:2265–2275.

Parker HG, VonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, Maslen CL, Acland GM, et al. 2009. An Expressed Fgf4 Retrogene Is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs. Science (80-) 325:995–998.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans - mobilized by the human LINE-1 protein machinery. Nucleic Acids Res 40:1666–1683.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–6.

Sayah DM, Sokolskaja E, Berthoux L, Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 430:569–573.

Schrider DR, Navarro FCP, Galante PAF, Parmigiani RB, Camargo AA, Hahn MW, Souza SJ de. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genet 9:e1003242.

Smith BF, Yue Y, Woods PR, Kornegay JN, Shin J-H, Williams RR, Duan D. 2011. An intronic LINE-1 element insertion in the dystrophin gene aborts dystrophin expression and results in Duchenne-like muscular dystrophy in the corgi breed. Lab Investig 91:216–231.

Watson CM, Camm N, Crinnion LA, Antanaviciute A, Adlard J, Markham AF, Carr

IM, Charlton R, Bonthron DT. 2017. Characterization and Genomic Localization of a SMAD4 Processed Pseudogene. J Mol Diagnostics 19:933–940.

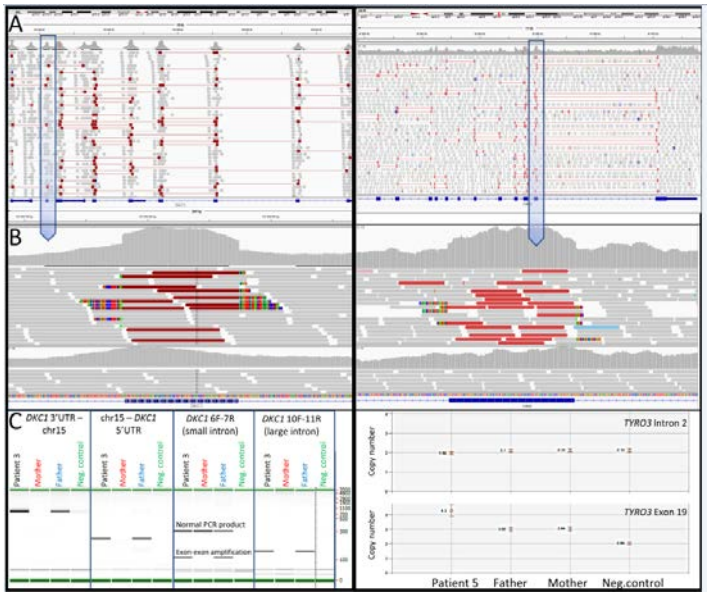
Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB. 2009. Mobile elements create structural variation: Analysis of a complete human genome. Genome Res 19:1516–1526.

Zhang Z, Harrison PM, Liu Y, Gerstein M. 2003. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13:2541–58.

Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, Ruzzo EK, Gumbs C, et al. 2012. Using ERDS to infer copy-number variants in high-coverage genomes. Am J Hum Genet 91:408–21.

Figure

Figure 1: Short read sequencing alignment visualization and confirmation of retrocopy. IGV v2.3 visualization at gene scale for patient 3 (left panel, *DKC1* gene analyzed by targeted sequencing) and 5 (right panel, *TYRO3* gene analyzed by genome sequencing) showing aberrant read pairs (colored in red) spanning entire introns. When zooming at exon level (B) the depth of coverage plot reveals an increase in exonic sequence with clear shifts at both exon-intron junctions. Aberrant read pairs are split reads with non-reference sequence in introns. (C) Confirmation techniques: exon-exon and *DKC1* retrocopy-insertion site junction fragment PCR amplification for the trio of patient 3 and results of ddPCR CNV testing for patient 5 confirming homozygosity of *TYRO3* retrocopy.



Table

Table 1: ¨Patients, tests and results summary. All genomic coordinates are expressed on GRCh37/hg19 human reference genome. ddPCR stands for droplet digital PCR

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8
Patient								
Sex	M	M	F	F	F	F	M	M
Age	69 years	6 months	10 years	19 years	55 years	3 weeks	Not known	12 years
Indication for referral	Attenuated adenomatous polyposis	Hydrocephalus	Intellectual disability	Intellectual disability	Early onset Alzheimer disease	De novo Balanced Translocation breakpoint characterization	Complex chromosomal rearrangement characterization	Osteochondrodysplasia
Methods								
Genetic test	11 genes panel sequencing (introns and exons)	hydrocoephalus panel sequencing (exons)	450 gene panel for intellectual disability (exons)	Exome sequencing	Genome sequencing	Genome sequencing	Genome sequencing	Exome sequencing
Targete	Agilent	Agilent	SeqCap	Agilent	Not	Not	Not	Agilent

d tehnology	SureSelect QXT	SureSelect QXT	EZ, Roche	SureSelect QXT	applicable	applicable	applicable	SureSelect XT
Sequencing machine	Illumina MiSeq	Illumina MiSeq	NextSeq. 500	Illumina HiSeq	BGISEQ-500	NextSeq. 500	HiSeq. 4000	HiSeq
Read length	2x150	2x150	2x75	2x150	2x50	2x101	2x151	2x150
Alignment program	BWA v0.7.12	BWA v0.7.12	BWA-MEM v0.7.10	BWA v0.7.12	BWA v0.7.12	BWA-MEM v0.7.10	BWA-MEM v0.7.10	BWA v0.7.12
SV Caller	GRIDSS v1.4.1	-	-	GRIDSS v1.4.1	Break Dancer v1.4.5	BreakDancer v1.4.5	BreakDancer v1.4.5	-
CNV caller	CANOES	CANOES	DeCovA	CANOES	ERDS v1.1	ERDS v1.1	ERDS v1.1	-
SNV Caller	GATK Haplotype Caller 3.8	GATK Haplotype Caller 3.8	GATK Haplotype Caller 3.8	GATK Haplotype Caller 3.8	GATK Haplotype Caller 3.8	-	-	GATK Haplotype Caller 3.8
	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7	Patient 8

Results

Retrocopied gene	*SMAD4*	*CCDC88C*	*DKC1*	*MTMR2*	*TYRO3*	*RCBTB1*	*C20orf27*	*RPL13*
Retrocopied transcript	ENST00000398417	ENST00000553403	ENST00000369550	ENST00000393223	ENST00000559066	ENST00000002586	ENST00000399672	ENST00000567815

Copy number gain	Yes (Strictly exonic)	Yes (Strictly exonic)	Yes (Strictly exonic)	Yes (Strictly exonic)	Yes (Strictly exonic)	Yes (Strictly exonic)	Yes (Strictly exonic)	-
Exonic SNV	NM_005359.5: c.869A>C not phased	No	No	No	NM_006293.3: c.2005G>T and c.1484-1G>T both phased to retrocopy	Not tested	Not tested	No
Confirmation technique	QMPSF, MLPA and PCR amplification	PCR amplification	PCR amplification	PCR amplification	ddPCR	PCR amplification	Not performed	PCR amplification
Inheritance	unknown (parents not available)	pat	pat	pat	biallelic	mat	unknown (parents not available)	biallelic
Insertion site								
Insertion cytoband	9q33.3	14q32.11	15q23	13q14.13	13q14.11	10p12.1	16p12.2	17p11.2
Insertion position [GRCh37]	chr9:127732713	chr2:24890767	chr15:67600596	chr13:47294695	chr13:44069850	chr10:28795699	chr16:22153620	chr17:17286666-17287378
Gene disruption	*SCAI* (intron 17)	*NCOA1* (Intron 5)	*IQCH* (intron 4)	*LRCH1* (Intron	*ENOX1* (Intron	No	*VWA3A* (intron	No

				15)	2)		24)	
CTCF site	No	No	No	No	No	No	No	No
TAD boundary	No	No	No	No	No	No	No	No
Same strand	No	No	No	No	Yes, in frame	No	Yes, not in frame	No